

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

This paper is made available online in accordance with publisher policies. Please scroll down to view the document itself. Please refer to the repository record for this item and our policy information available from the repository home page for further information.

To see the final version of this paper please visit the publisher's website. Access to the published version may require a subscription.

Author(s): Alix, Karine; Joets, Johann; Ryder, Carol D.; Moore, Jay; Barker, Guy C.; Bailey, John P.; King, Graham J.; Pat Heslop-Harrison, John S

Article Title: The CACTA transposon Bot1 played a major role in Brassica genome divergence and gene proliferation

Year of publication: 2008

Link to published version: <http://dx.doi.org/10.1111/j.1365-313X.2008.03660.x>

Publisher statement: The definitive version is available at www.blackwell-synergy.com

The CACTA transposon *Bot1* played a major role in *Brassica* genome divergence and gene proliferation

Karine ALIX^{1*}, Johann JOETS¹, Carol D. RYDER², Jay MOORE², Guy C. BARKER², John P. BAILEY³, Graham J. KING⁴ and J. S. (Pat) HESLOP-HARRISON³

¹UMR de Génétique Végétale INRA/Univ Paris-Sud/CNRS/AgroParisTech, Ferme du Moulon, F-91190 Gif-sur-Yvette, France

²Warwick HRI, Wellesbourne, Warwick CV35 9EF, UK

³Department of Biology, University of Leicester, Leicester LE1 7RH, UK

⁴Rothamsted Research, Harpenden, Hertfordshire, AL5 2QJ, UK

*author to whom correspondence should be addressed

e-mail: alix@moulon.inra.fr, phone: +33/(0)1 69 33 23 72, fax: +33/(0)1 69 33 23 40

Running title: Role of CACTA *Bot1* in *Brassica* genome evolution

Keywords: *En/Spm* transposable element, *Brassica oleracea*, *Brassica rapa*, genome evolution, *SLL3* gene, *S* locus

The TE sequences reported in this paper have been deposited in the EMBL database under accessions nos. AM888354-AM888369; the three *B. oleracea* BAC sequences have been deposited under accessions nos. EU642504-EU642506.

Word count: 6,938

Submit to The Plant Journal

SUMMARY

We isolated and characterized a *Brassica* C genome-specific CACTA element, designated *Bot1* (*B. oleracea* *transposon* *1*). Analysing phylogenetic relationships, copy numbers and sequence similarity of *Bot1* and *Bot1*-analogues in *B. oleracea* (C genome) vs. *B. rapa* (A genome), we concluded that *Bot1* has encountered several rounds of amplification in the *oleracea* genome only, playing a major role in the recent *rapa* and *oleracea* genome divergence. We performed *in silico* analyses of the genomic organization and internal structure of *Bot1* and establish which segment of *Bot1* is C genome specific. Our work represents the first report of a fully characterised *Brassica* repetitive sequence which can distinguish the *Brassica* A and C chromosomes in the allotetraploid *B. napus*, by fluorescent *in situ* hybridization. We demonstrated that *Bot1* carries a host *S* locus-associated *SLL3* gene copy. We speculate that *Bot1* was involved in the proliferation of *SLL3* around the *Brassica* genome. The present study reinforces the assumption that transposons are major drivers of genome and gene evolution in higher plants.

INTRODUCTION

Repetitive DNA sequence motifs represent a large fraction of the eukaryotic genome; they account for at least 50% of the human genome (IHGSC, 2001), and in some plants they can constitute up to 80% of the genome (Meyers *et al.*, 2001; Vicient *et al.*, 2001). Repetitive sequences can be divided into two broad groups: those that are tandemly repeated and form large blocks in the genome, and those that have a dispersed genomic distribution and include transposable elements (TEs) (for review Heslop-Harrison, 2000). Transposable, or mobile, elements are divided into two main classes (Wicker *et al.*, 2007), according to the presence or absence of an RNA transposition intermediate: class I elements or retrotransposons move via an RNA intermediate, which is reverse transcribed prior to its reintegration into the genome; class II transposons directly move as DNA elements, including transposons that copy themselves for insertion (subclass 2) and the others that leave the donor site to reintegrate elsewhere in the genome (subclass 1).

It is now assumed that repetitive elements, especially TEs, are major drivers of genome and gene evolution (for reviews: Bennetzen, 2005; Casacuberta and Santiago, 2003; Fedoroff, 2000; Kazazian, 2004). In plants, nuclear genome sizes vary tremendously between species; while polyploidy (a key evolutionary process in plants; Gaeta *et al.*, 2007; Wendel, 2000) is often associated with rapid and dramatic changes in sizes of the constituent genomes, it has been shown that most genome size variability is associated with differences in repetitive DNA content and is mainly ascribed to differential amplification of TEs (Hawkins *et al.*, 2006). There is also a possible higher order chromosome structural role for repeats, related to packaging of genes, DNA methylation or histone modifications and genetic regulation, with consequential effects

on rate of recombination: large-scale genome rearrangements are commonly found near or at regions enriched for repeated DNA (Eichler and Sankoff, 2003).

The role of transposons in gene evolution in plants has been demonstrated only recently. Class II transposons preferentially target gene-rich regions for insertion, as observed in *Arabidopsis*, rice and maize (for reviews, Feschotte *et al.*, 2002 and Walbot and Petrov, 2001). Studies on flowering plants provided evidence for the involvement of transposons in generation and evolution of genes by mechanisms of transposon capture and exon shuffling (for reviews, Bennetzen, 2005 and Morgante, 2006). Examples of transposons mediating gene movements were first represented by Pack-MULEs (*Mutator*-like transposable elements) found in maize (Talbert and Chandler, 1988), *Arabidopsis* (Yu *et al.*, 2000) and rice (Turcotte *et al.*, 2001). In rice, it has been estimated that 3,000 Pack-MULEs contain genic fragments derived from more than 1,000 different cellular genes, with some of the captured genes being transcribed and potentially functional (Jiang *et al.*, 2004). *Helitrons*, which transpose by rolling-circle replication (for review Kapitonov and Jurka, 2007), were discovered by computer-based analysis of genomic sequences from *Arabidopsis*, *Caenorhabditis elegans* and rice, and were reported to have recruited host genes, and multiplied them in the host genomes (Kapitonov and Jurka, 2001), seen also in maize (Brunner *et al.*, 2005; Morgante *et al.*, 2005). Elements of another superfamily of subclass 1 transposons, called CACTA or *En/Spm* (referred to the first element described in maize, *Enhancer/Suppressor-mutator* – Peterson, 1953), were also able to capture host cellular genic fragments (Kawasaki and Nitasaka, 2004; Rocarro *et al.*, 2005; Zabala and Vodkin, 2005) which were retained as functional coding sequences (Zabala and Vodkin, 2007). According to these different studies, transposons have the potential to amplify genes across the genome and

to create novel genes by capturing and rearranging genic sequences, which thus represents an important mechanism for the evolution of genes in higher plants and can lead to rapid diversification of genome lineages during speciation.

Access to the complete genomic sequence of *Arabidopsis thaliana* demonstrated that *Arabidopsis* harbours all of the TE types found in larger plant genomes (AGI, 2000), and a recent comparative analysis based on bioinformatics showed that *A. thaliana* and *Brassica oleracea* (*Brassica* genome CC) share largely the same collection of TEs (but in differing proportions, the number of elements of each type of TEs being greater in *B. oleracea*) (Zhang and Wessler, 2004), in accordance with the high level of sequence conservation between the two species which diverged from a common ancestor 15-20 million years ago (Yang *et al.*, 1999). Up to now, only a very small fraction of the *Brassica* TEs has been studied and analysed at the molecular level; indeed, efforts were focused on the isolation, sequence characterization and molecular evolution mainly of class I TEs including LTR-retrotransposons and LINEs (Alix *et al.*, 2005; Alix and Heslop-Harrison, 2004), SINEs (*Sl*: Deragon *et al.*, 1994; *BoS*: Zhang and Wessler, 2005), and recently TRIM (Yang *et al.*, 2007). Interestingly, it has been estimated that the most abundant class II superfamily in *B. oleracea* is CACTA (Zhang and Wessler, 2004). The designation ‘CACTA’ refers to the flanking terminal inverted repeats (TIRs), normally 10 to 28bp long, which terminate in a conserved 5’-CACTA-3’ motif. CACTA elements also have characteristic sub-terminal repetitive regions of 10 to 20bp units which are repeated in direct and inverted orientations. The *En/Spm* transposable element system of maize was the first CACTA element that was isolated and characterized at the molecular level (Pereira *et al.*, 1986). Internal sequences of CACTA

elements are highly variable, and it is difficult to identify CACTA transposons from only the analysis of sequence similarity; therefore, most CACTA transposons were found because of the presence of a transposase-like protein and a terminal CACTA motif. CACTA elements seem to be found mostly in plants (DeMarco *et al.*, 2006; Wicker *et al.*, 2003), and several autonomous and active CACTA (*En/Spm*) transposons were isolated and characterized in various plant species including *Tam1* in snapdragon (Nacken *et al.*, 1991), *Tdc1* in carrot (Itoh *et al.*, 2003; Ozeki *et al.*, 1997), *Tpn1* in Japanese morning glory (Inagaki *et al.*, 1994), *Ps1* in petunia (Snowden and Napoli, 1998), *Rim2* in rice (He *et al.*, 2000), *CAC1* in *Arabidopsis* (Miura *et al.*, 2001) and in *Beta vulgaris* (Jacobs *et al.*, 2006).

In the present work, we isolated and characterized a *Brassica* C genome specific CACTA transposon, designated *Bot1* (*Brassica oleracea* *transposon 1*). We assessed the genome-specificity of this *B. oleracea* class II transposon by carrying out sequence, molecular and cytogenetic analyses on representatives of six diploid and allotetraploid *Brassica* species; our results show that *Bot1* has proliferated within the C genome contributing to *Brassica* genome divergence. Moreover, characterising *Bot1*, we have identified the domain which appears to be C genome specific and a new example of a transposon capturing a host gene, with the systematic finding of the *S* locus-associated *SLL3* gene copy inside the CACTA *Bot1*; we speculate that *Bot1* was involved in the multiplication of *SLL3* across the *Brassica* genome. Thus this study confirms that transposons are major drivers of genome and gene evolution in higher plants.

RESULTS

Isolation of *Bot1*, a *Brassica* C genome-specific transposon

We carried out PCR using the degenerate primer pair BEL1MF/BEL2MR (originally designed to amplify LINE-related retrotransposons; Kubis *et al.*, 2003) with genomic DNA from the three diploid *Brassica* species, *B. rapa* (*Brassica* genome AA), *B. nigra* (BB), and *B. oleracea* (CC), and the three allotetraploid species, *B. napus* (AACC), *B. juncea* (AABB) and *B. carinata* (BBCC) of U's triangle (1935) (plant material is listed in Table 1). A 410bp fragment was amplified from all six species and shown to represent LINEs (Alix and Heslop-Harrison, 2004). In contrast, and fortuitously, a 1kb fragment was amplified only in *Brassica* species including the C genome: *B. oleracea* (CC), *B. napus* (AACC) and *B. carinata* (BBCC). The 1kb fragment was tested on a set of 24 accessions and was observed in 9 *B. oleracea* accessions, 9 *B. napus*, and 1 *B. carinata*, but not seen in 2 *B. rapa*, 1 *B. nigra*, or 2 *B. juncea* (a subset of the results obtained is given in Figure 1a), and further tests showed the 1kb fragment was also amplified with the single primer BEL1MF (Figure 1a; BEL2MR alone gave no products).

We cloned representatives of the 1kb-band from the three species possessing the C genome (Table 1). To examine the genomic distribution of the amplicon, clone Bo6L1-15 (from *B. oleracea*, see Methods for nomenclature, EMBL accession #AM888359) was used for Southern hybridization to *Eco*RI-digests of genomic DNA from the six *Brassica* species; hybridization signals were only detected in *Brassica* species with the C genome, with stronger hybridization to *B. oleracea* (Figure 1b). Analysis of the 16 sequences from various accessions of diploids and allopolyploids with the C genome (EMBL accession nos. AM888354-AM888369) showed ten were

around 1010bp long; five others had a single deletion of 29bp relative to these, while one had an additional 69bp deletion (Figure S1). Fifteen of the sequences contained a 312bp ORF encoding a protein of 103 amino acids (Bn14L1-25 had a SNP, TGA → GGA). Sequences were conserved within coding (average of 81% at DNA level) and non-coding (67%) regions.

BLASTN comparisons using Bo6L1-15 (1010bp) as a query sequence identified homologous fragments in three BAC sequences from a *B. oleracea* BAC library (the ‘BoB’ library from Warwick-HRI see Alix *et al.* (2005) which correspond to a triplicated region within the *B. oleracea* genome (C. Ryder, unpublished data): BoB028L01 EMBL accession number EU642504, BoB048N13 #EU642506, and BoB029L16 #EU642505 with respectively 63%, 92%, and 96% similarity to Bo6L1-15. BLAST comparisons also indicated that Bo6L1-15 was homologous to fragments in three *B. oleracea* sequence contigs (contigs A, G, and B: EMBL accessions #AC183495, AC183492 and AC183493; Town *et al.*, 2006), again corresponding to two sets of paralogous triplicated sequences within the *B. oleracea* genome (O’Neill and Bancroft, 2000). We obtained other significant sequence identities with a single *B. napus* clone corresponding to the non-coding part of the *S* locus region (EMBL accession #AJ245479; Cui *et al.*, 1999) and eight *B. rapa* BAC sequences (AC189496, AC189360, AC189655, AC189480, AC189314, AC189341, AC189446, AC189258). The conserved 103aa-long ORF exhibited homologies to multiple plant *En/Spm*-like transposases (from *Arabidopsis*: AAF06087, AAG50553; from carrot: AB070979; or from *Medicago*: ABE78938). We chose the designation *Bot1*, ‘*Brassica oleracea* *transposon 1*’, for this *Brassica* C genome-specific element. BLAST searches with the ORF of the *Bot1* representative Bo6L1-15 as a query sequence against the *Brassica* EST

database revealed strong homologies between the *Bot1* coding sequence and one *B. oleracea* EST (AM394037), the *B. napus* EST CN729093 and a set of forty two *B. napus* ESTs (UniGene Bna.6539). Interestingly, the *B. oleracea* EST AM394037 corresponds to a drought-stress induced EST in the database (Paniwnyk Z., Pink D., Akehurst J., Buchanan-Wollaston V., unpublished work), and the *B. napus* EST CN729093 has been identified at early stages of seed development under abiotic stress conditions (Georges F., unpublished).

We performed phylogenetic analyses on the 16 *Bot1* ORF sequences we isolated and 13 other ORF sequences of *Bot1* from *B. oleracea* or of *Bot1*-related elements from *B. napus* and *B. rapa* obtained from the three 'BoB' BAC sequences or mined from the database (including those published by Town *et al.*, 2006; see methods). Topologies of trees from the maximum parsimony (MP), maximum likelihood (ML; Figure 2) and neighbour-joining (NJ) analyses were similar, showing two main clades supported by a bootstrap value of 100%, with *Bot1* constituting a C genome-specific phylogenetic clade (Figure 2). The *Bot1* clade is marked by unresolved polytomies (rake-like branching) suggesting a low rate of sequence divergence, in accordance with the extensive identity we observed among the ORFs of *Bot1* (see above). In contrast, the second major clade which tends to be *Brassica* A genome-specific shows a well-resolved phylogeny with all branches supported by high bootstrap values (Figure 2) suggesting a relatively high rate of sequence divergence.

***Bot1* allows *Brassica* A and C chromosomes to be distinguished**

In situ hybridization on metaphase plates from *B. oleracea* and *B. napus* using clone Bo6L1-15 as a probe showed specificity of *Bot1* and its physical distribution along the chromosomes. The probe hybridized to multiple sites dispersed along the full length of

all 18 chromosomes in *B. oleracea* and 18 of the 38 chromosomes in *B. napus* (Figure 3), with gaps around most centromeres, some chromosome arms with less signal, and a few stronger sites seen as dots on both chromatids. Thus the elements were abundant and widely dispersed throughout the C genome. A few minor hybridization signals were detected on the 20 A-genome chromosomes, supporting the presence of *Bot1*-analogues in the genome of *B. rapa*, in accordance with the results obtained from the BLAST similarity search.

***Bot1* belongs to a *Brassica* family of CACTA elements**

We subjected the three ‘BoB’ BAC sequences including CACTA homologies to dot plot analyses in order to identify exact or degenerate repeats (see methods) and CACTA motifs (Figure 4a). The analysis delimited a single full-length CACTA element for each BAC analysed; size and position of the CACTA within each BAC are thus provided in Figure 4b. TIRs of 15 and 17bp were identified in BoB028L01 (CACTACAAGAAAACA) and BoB029L16 (CACTACAAGAAAACAGC), while the third TIR, in BoB048N13, was 64bp long (CACTACAAGAAAACAGCGATATTCTGACGGACATTCCGACGGAAAATGAAATCCTCGGAATATA). The first 15 nucleotides from each TIR were conserved across the three BAC sequences; sub-terminal repeats were also found (Figure 4a, 4b). In addition, ‘CTA’ 3bp-target site duplication (TSD) which is characteristic for the CACTA superfamily (Wicker *et al.*, 2007) was identified for each *Bot1* insertion. Interestingly, the 5’ and 3’ flanking sequences of *Bot1* were AT-rich with an average GC content of less than 30 % (Figure 4c), 41% GC content being considered average for the C genome (J. Moore, unpublished data). These three elements showed similar overall structures, distinguished by sequence insertions and rearrangements (Figure 5).

CDS with identities to putative retroelement *pol* polyproteins (28% identity with AAM15254 from Arabidopsis) were identified in *Bot1-2* and *Bot1-3* (Figure 4b), with approximately 80bp overlapping the Bo6L1-15 sequence we used as a probe (Figure 5). The presence of stop-codons in the internal part of the CDS which is relatively small compared to its homologue in Arabidopsis, suggests that this CDS is a pseudogene. Considering the CACTA domain matching the Bo6L1-15 sequence (Figure 4b), it seems that the use of a degenerate primer originally designed to target the reverse transcriptase of LINEs (a gene that is included in the *pol* region of retrotransposons), led to non-specific amplifications of a short section of a gene coding for TE-related proteins, namely a transposase and a *pol* protein-like in the present case. Fortuitously, as part of this non-specific amplicon, this degenerate primer goes on to amplify one of the C genome-specific parts of *Bot1* (Figure 5). Database BLAST comparisons identified full-length CACTA elements from both *B. rapa* and *B. napus* BAC sequences; locations of these elements within each BAC, as well as size and TIR sequences are reported in Table 2. The overall structure of these elements is similar to those characterized in *B. oleracea*; however, when we compared the three *B. oleracea Bot1* representatives to the CACTA elements identified in the genomes of *B. napus* and *B. rapa*, we observed that all *Bot1* copies contain *B. oleracea*-specific sequence fragments (orange domains in Figure 5); for instance, within the region corresponding to the Bo6L1-15 fragment, two thirds of the nucleotide sequence is specific to *B. oleracea* for two of the BACs (Figure 5). Using the complete 10.9kb of *Bot1-2* against *B. oleracea* WGS and *B. rapa* BAC end sequences the number of BLAST alignments to each genome was investigated (Figure S2). This analysis again highlighted a C genome proliferation but also revealed

a 2.7kb region in which we found no *B. rapa* alignments. This region appears to be C genome specific.

The transposase sequence of CACTA elements is not well defined; we looked for potential active transposases in the *Brassica* CACTA elements. The *B. rapa* BAC sequence AC189480 contains a complete and intact transposase-encoding gene of 3,363bp (1,120aa) which belongs to the Tnp2-like family, while the homologous gene of *B. oleracea* is interrupted by insertions (orange domains) or deletions (*Bot1-3*) which appear C genome-specific (Figure 6). We found that the 103aa-long putative transposase-encoding ORF that was first identified in the Bo6L1-15 fragment (see above) corresponds to the C-terminal part of the transposase (Figure 6), and can be classified as a pseudogene. We conclude that the different ESTs mined from expressed sequence database as similar to the Bo6L1-15 ORF may originate from another complete version of the transposase gene present elsewhere in the *B. oleracea* genome. In addition, the coding sequence of each putative transposase of the three *B. oleracea* CACTA elements contains several in-frame stop codons. As we could not identify any other gene with a potential transposase function in *Bot1*, we believe that the *Bot1* representatives we analysed do not possess any functional and still active transposase.

***Bot1* carries a host *SLL3* gene copy**

Notably, one copy of the *Brassica* *S* locus-associated *SLL3* gene, first described in *B. napus* by Cui *et al.* (1999), was found in all the *Bot1*-related CACTA elements we characterized in *B. oleracea* and *B. napus/B. rapa* (Figure 4b and data not shown). Likewise, the three *B. oleracea* contig fragments (Town *et al.*, 2006) we identified as possessing sequences highly similar to the *Bot1* clone Bo6L1-15 by BLAST similarity search also carry a *SLL3* gene downstream from the *Bot1* matching sequence. The

general structure of the *SLL3* copies was highly conserved in the various CACTA elements we analysed, including 5 exons and 4 introns as previously described by Cui et al. (1999), except for the first intron sequence: the corresponding genes identified in *B. oleracea* CACTA elements exhibit an alternative shorter intron 1 sequence compared to the published *B. napus* gene, and *SLL3* copies included in *B. rapa* *Bot-1*-like CACTA elements contain a short insertion in the first intron. The very low GC content of *SLL3* intron sequences was intriguing. We found five different PlantGDB-assembled Unique Transcripts (PUT) highly similar to *SLL3* (90 to 97% identity) which overlap together 80% of the gene, although no PUT supported the four intron – five exon structure suggested by Cui et al. (1999).

***Bot1* and *SLL3* have proliferated in the *B. oleracea* genome**

We estimated the copy number of *Bot1* in the *B. oleracea* genome by examining the presence of this element in a portion of the ‘BoB’ BAC library representing 2.8× genome coverage (Alix et al., 2005). A total of 2,110 BACs hybridized to the probe Bo6L1-15. Assuming the *B. oleracea* genome is fully represented in the BAC library and there is no clustering of copies of the element on single BACs, we predict 755 genomic insertion sites for *Bot1*, per haploid genome. Analysis of three different *Bot1* sequences (Bo6L1-15, Bo1L1-04 and Bo8L1-11) against *B. oleracea* whole genome shotgun (WGS) sequences gave estimates of between 1,420 and 1,629 copies in the *B. oleracea* genome (Table 3); the three sequences hit the same set of *B. oleracea* WGS sequences, demonstrating that we isolated representatives of a single family of CACTA transposons. For the two clones Bo6L1-15 and Bo1L1-04, we performed BLAST comparisons using first the transposase-CDS only and then the full 1kb-sequence: we obtained closely similar copy number estimates in both cases, demonstrating that the *pol*

polyprotein retroelement present in *Bot1* (Figure 4b) is not found elsewhere in the genome and now constitutes a specific part of *Bot1*. The two-fold higher estimates of genome copy number from WGS analysis compared to BAC library hybridization suggests some clustering of *Bot1*. Comparison of results obtained at different E value stringencies reveal only small differences; this is indicative of a relatively low rate of sequence divergence for this CACTA element. We performed the same analysis using the complete Arabidopsis genomic sequence and *B. rapa* BAC sequences (BAC ends (BES) & whole BACs): no alignments to any sequence in the Arabidopsis genome were found and only approx. 30 in *B. rapa* (Table 3). To evaluate the dynamics of *Bot1* in the *Brassica* genomes, we used 266 *Bot1* ORF sequences mined from *B. oleracea* WGS (plus the ORF of Bo6L1-15) and 14 *Bot1*-like ORFs from *B. rapa* BAC sequences in Genbank to calculate a median network phylogeny (Figure 7) and a phylogenetic tree (Figure S3). For the median network phylogeny we found that tree topologies from the maximum parsimony (MP), maximum likelihood (ML) and neighbour-joining (NJ) analyses were identical in almost all aspects. The median network phylogeny distinguishes of a group of 12 *B. rapa* copies constituting a *B. rapa*-specific clade which is monophyletic. The other A genome copies are closer to some copies of the C genome. This same trend is observed in the phylogenetic tree (Figure S3). Such a topology suggests that only a few copies of the CACTA transposon were present in the common ancestor, while the star contractions in the median network phylogeny suggest two or three main expansions in the C genome. The phylogeny contains considerable reticulation (i.e. not a straightforward tree) indicating that there may well have been recombination through conversion, complicating the phylogenetic relationships. The reticulations are independent between the two genomes, indicating that such

recombination would be expected to have occurred since speciation. Thus Figure 7 and Figure S3 echo the trend observed in Figure 2, produced using a smaller data set, with one particular *B. oleracea* clade marked by unresolved polytomy with a typical rake-like branching pattern, indicating a low rate of sequence divergence.

Analysis of the *SLL3* gene sequence against *B. oleracea* WGS and *B. rapa* BAC and BAC-end sequences gave estimates of between 2,490 and 4,240 copies in the *B. oleracea* genome and only 395 to 910 copies in the *B. rapa* genome (Table 3), with no evidence for copies in the Arabidopsis genome. Parallel copy number estimates suggest that *SLL3* has been more highly proliferated within the C genome. It is noteworthy that this high copy number of *SLL3* in the C genome contrasts strongly with the retrieval of only five *B. napus* PUT from the EST database.

DISCUSSION

In this study, we isolated and characterized a new family of *Brassica* CACTA elements (*En/Spm* transposons), with representatives from *B. oleracea*, named *Bot1*. We demonstrated *Brassica* C genome proliferation and identified a 2.7kb region of C genome specificity. Full-length but defective elements averaging 10kb in length were identified by BAC library screening. BLASTN analysis against *B. oleracea* WGS showed that about 1,500 copies of *Bot1* are present in the *B. oleracea* haploid genome, representing some 2.3% of the genome, or more than 10% of the transposable elements (Zhang and Wessler (2004) estimate that TEs constitute 20% of the *B. oleracea* genome). It is noteworthy that *Bot1* represents the first *B. oleracea*-specific dispersed repeated sequence which has been isolated.

***Bot1* played a role in the recent *Brassica* genome divergence and shows evidence for genomic homogenization events**

The evolutionary history of TEs is important to study because of the effect of these elements on the structure of the plant genome and its evolution. Our analysis of the *Bot1* element in BACs and contigs showed its presence in homologous sites in triplicated regions of *B. oleracea*, suggesting that an element was present in the ancestral *Brassicaceae* genome which then underwent triplication of its genome (Lysak *et al.*, 2005). Our data also identified, at low copy number, a *Bot1* family-related sequence in the *B. rapa* genome. Hence we suggest that evolutionarily recent homogenization, based on an existing element variant, has allowed replacement of other, older, elements, perhaps combined with some amplification and new dispersal, giving the characteristic dispersed and high copy number C-genome-specific element that we observe. Such a mechanism has been suggested for tandemly repeated satellite DNA sequences in *Drosophila* (Strachan *et al.*, 1985) with increasing evidence coming from sequence and cytogenetic analysis (Kuhn *et al.*, 2008).

Another notable feature of *Bot1* is the low sequence divergence, a result in accordance with Zhang and Wessler (2004) who performed phylogenetic analyses on CACTA-like elements in *Arabidopsis* and *B. oleracea*, and demonstrated a high intra-family sequence identity for the *B. oleracea* elements. The combination of PCR, bioinformatics, DNA and chromosomal hybridizations used here all show low divergence within the family, and high divergence from other families in the A and B genome species, demonstrating that the result is not a consequence of selective sampling. Our phylogenetic tree with 280 *Bot1* sequences indicates expansion through two or three rounds of amplification (Figure 7). High sequence identity is normally

taken to indicate that an element has amplified recently; our result suggests that *Bot1* has amplified through homogenization and replacement of other elements in the C genome (from a low and undetectable ancestral *Bot1* copy which is present in the common *Brassica* progenitor), after divergence from the A genome approximately 3.75 million years ago (Inaba and Nishio, 2002).

***Bot1* is a non-autonomous but apparently still active CACTA transposon**

Bot1 and related *Brassica* elements display all characteristics of CACTA transposons, including (i) the presence of TIRs of 15-17 bp in size containing the consensus sequence CACTACAAGAAA (Kawasaki and Nitasaka, 2004) and sub-terminal regions consisting of both direct and inverted repeats, (ii) a 3bp-TSD, (iii) ORFs showing sequence identities to transposases of previously characterized plant *En/Spm* transposons, and (iv) preference for insertion in AT-rich regions (Le *et al.*, 2000; Nacken *et al.*, 1991; Wang *et al.*, 2003). One element analysed, *Bot1-3*, harbours unexpectedly long TIRs of 64bp; this could reflect analogies with MULEs (*Mutator*-like elements) as transposons with long TIRs (e.g. *Mu* in maize). The complete *Bot1* representatives we identified in *B. oleracea* were defective and thus non-autonomous, although an apparently complete *Bot1*-like transposase gene was found in *B. rapa*, and enzymes from such elements could trans-activate the movement of defective elements. BLAST searches for *Bot1* in *B. oleracea* and *B. napus* EST database showed that some *Bot1* copies are transcribed into polyadenylated RNA, demonstrating the existence of *Bot1* transcripts. It is noteworthy that the *B. oleracea Bot1* homologue EST was isolated during the study of the effect of pre-harvest drought stress on gene expression in post-harvest broccoli (Paniwnyk *et al.*, unpublished work) and that *B. napus* ESTs similar to

Bot1 were also isolated during *B. napus* seed development under abiotic stress conditions (Georges *et al.*, unpublished work). Stress, including that from interspecific hybridization, is now accepted to lead to transposable element activity (Grandbastien, 1998) but the impact on turnover of transposon sequences is still not clear.

C-genome specificity of *Bot1*

The results shown here demonstrate the value of *Bot1* as a genetic marker for the *Brassica* C genome. Up to now, A or C genome-specific markers were mostly derived from diverged sequences and consisted of various nucleotide substitutions (e.g. SINE S1 – Lenoir *et al.*, 1997; various RFLP markers – Gaeta *et al.*, 2007). In an analysis of the Brassicas in the U triangle, PCR with the single primer BEL1MF gives a product only from species including the C genome, and hence could be used to identify these species, including any C genome introgression, in DNA extracts from plant bulks. *In situ* hybridization demonstrated that the *Bot1* sequence was abundant throughout the C genome and present on all chromosomes, and that hybridization sites were restricted to the C chromosomes. Thus, the sequence can distinguish A and C chromosomes efficiently in the allotetraploid species *B. napus*. There are no previous reports of such dispersed C genome-specific repeats. Gene sequence analysis shows that the A and C genome species *B. rapa* and *B. oleracea* are more related to each other than to the B genome *B. nigra* (Lysak *et al.*, 2005). As a consequence, classical genomic *in situ* hybridization (GISH) allows the B genome chromosomes to be distinguished from the A or C genome chromosomes in natural allotetraploids, while it fails to differentiate efficiently the A and C genomes in *B. napus* because of extensive cross-hybridization between the two genomes (Snowdon *et al.*, 1997) and stronger labelling of the centromeric regions and their genome-specific satellite repeats. The *Bot1* sequence

distribution along the chromosomes, with exclusion from major satellite repeats around the centromere, and some variation in hybridization strength along the arms, was similar to that found for an *En/Spm* element in *Beta vulgaris* (Jacobs *et al.*, 2006). The CACTA transposon *Bot1* appears to be of great value for studies on the evolution of the *Brassica* genome, for physical mapping or for monitoring introgression in oilseed rape thanks to the detection of intergenomic recombination. The *B. oleracea* BAC BoB014O06, that was shown to contain *Bot1* according to our results from BAC filter hybridization using Bo6L1-15 as a probe, has already been used successfully to analyse recombination in AAC triploid hybrids (Leflon *et al.*, 2006) and to differentiate allosyndesis from autosyndesis in *B. napus* haploids (Nicolas *et al.*, 2007); this demonstrates the high efficiency of *Bot1* in distinguishing A and C *Brassica* chromosomes. In particular, in a context of reducing gene flow from genetically modified oilseed rape to its wild related counterpart represented mostly by *B. rapa*, *Bot1* will be very helpful in monitoring transgene introgression specifically into C-genome chromosomes.

***Bot1* was involved in the proliferation of the host *SLL3* gene**

Bot1 and other *Brassica Bot1*-analogues contain the *SLL3* gene from the *Brassica S* locus, inside the CACTA element. We note that an 8kb *Bot1*-analogue present in *B. rapa* was previously identified when characterizing the genomic organization of the *S* locus (Suzuki *et al.*, 1999). Sporophytic self-incompatibility (SI) is controlled by a single Mendelian genetic locus, the *S* locus (Bateman, 1955). Genomic characterization of this complex locus has demonstrated the occurrence of more than ten genes including three major highly polymorphic genes responsible for SI: the two genes *SLG* and *SRK* encode the female determinant of SI and *SP11/SCR* determines the *S* haplotype specificity of pollen (Shiba *et al.*, 2001). Cui *et al.* (1999) identified the *S* locus-linked

gene 3, *SLL3*; this gene encodes a putative secreted small peptide but its functional role is still unknown, as no functional evidence could be obtained due to its multiple-copy nature and the difficulty in studying its expression pattern (Cui *et al.*, 1999). Our study has confirmed the multiple-copy nature of *SLL3*, and strongly suggests that the copy number of *SLL3* is directly linked to the level of amplification of *Bot1*: indeed, we estimated 395-910 *SLL3* copies in the *B. rapa* genome in which *Bot1*-like sequences were slightly amplified while more than 3,000 *SLL3* copies were estimated in the *B. oleracea* genome where *Bot1* has proliferated. Such data suggest that the proliferation of *SLL3* across the *B. oleracea* genome is related to the proliferation of *Bot1*. The intron/exon structure of all the *SLL3* copies we analysed demonstrated that *SLL3* was captured from the eukaryotic host genome, eliminating the hypothesis that *SLL3* could be a gene necessary for transposition of the CACTA element which potentially originated from prokaryotic genome. Further study of the *SLL3* gene may enable an estimate to be made on the date of incorporation into the CACTA element. The hypothesis that most *SLL3* copies are pseudogenes due to mutation accumulation, could be tested along with the fact that the high *SLL3* gene copy number seems infrequently transcribed in the *Brassica* genome.

Conclusion

We report here the first example of a transposon mediating gene movement and proliferation across the *Brassica* genome. *Bot1* thus increases the number of transposons which have the potential to capture host genes, such as the CACTA elements identified in soybean (Zabala and Vodkin, 2005 and 2007), the Pack-MULEs characterized in maize (Talbert and Chandler, 1988), *Arabidopsis* (Yu *et al.*, 2000) and rice (Jiang *et al.*, 2004; Turcotte *et al.*, 2001), and the more recently discovered

helitrons in maize (for review Kapitonov and Jurka, 2007 and references therein). Moreover, we demonstrated that *Bot1* played a major role in the recent A and C *Brassica* genome divergence, supporting transposon amplification and homogenization as one of the major mechanisms of plant genome expansion and evolution (Bennetzen, 2005). The present study reinforces the view that the so-called “junk DNA” is a major driver of genome and gene evolution and diversification, demonstrating the need and importance of analysing repetitive sequences and in particularly TEs, to fully depict the evolution of the plant genome.

EXPERIMENTAL PROCEDURES

PCR amplification and cloning

Table 1 shows the identification and origin of the *Brassica* accessions used in the present study. Genomic DNAs were extracted from fresh leaves of three-week old plants following a standard CTAB protocol. For PCR amplifications, degenerate oligonucleotide primers were used from conserved domains flanking parts of LINE reverse transcriptase genes: BEL1MF = 5’RVN-RAN-TTY-CGN-CCN-ATH-AG3’ (45% degeneracy, T_m=55°C) and BEL2MR = 5’GAC-ARR-GGR-TCC-CCC-TGN-CK3’ (25% degeneracy, T_m=67°C) (Kubis *et al.*, 2003). PCR reactions were carried out according to Kubis *et al.* (1998) with an annealing temperature of 48°C for 1 min. PCR products of approximately 1kb in length were cloned (pGEM®-T, Promega) and transformed in *Escherichia coli* strain XL1-Blue. Clone names were established as follows: Bo for *B. oleracea*, Bn for *B. napus*, Bc for *B. carinata*, then the plant accession code number as given in Table 1 followed by ‘L1’ with ‘L’ for LINE primer

and '1' to indicate the 1kb band, a hyphen and a serial number to identify individual bacterial clones.

Southern blot hybridization

Five micrograms of total genomic DNA were restricted with the endonuclease *EcoRI* (Gibco BRL), separated in 1.2 % agarose gels and transferred onto positively charged nylon membranes (Roche) following standard procedures. The Bo6L1-15 probe was labelled with [α -³²P]dCTP by random priming using the Amersham Pharmacia Biotech Oligolabelling kit and purified on Sephadex columns. Southern blot hybridization was performed at high stringency as described previously (Alix *et al.*, 1998). Autoradiograph exposure of the Southern blot was 2 days.

Chromosomal *in situ* hybridization

Chromosome spreads were prepared from root tips of *B. napus* cv. 'Yudal' and 'Drakkar'. Methods for chromosome spreads and *in situ* hybridization essentially followed Alix *et al.* (2005) and Schwarzacher and Heslop-Harrison (2000). Bo6L1-15 was labelled with biotin-11-dUTP or digoxigenin-16-dUTP for use as a probe, and the most stringent wash was carried out in 20% formamide, 0.2× SSC at 42 °C, allowing labelled probe sequences with more than 85% homology to the chromosomes to remain hybridized. Images were captured on a Zeiss fluorescence microscope and processed with Adobe Photoshop using only functions which affect the whole image equally, except filling in some out-of-frame dark areas.

Sequence and phylogenetic analyses

Sixteen *Bot1* clones from seven accessions of the three *Brassica* species including the C genome, namely *B. oleracea* (C), *B. napus* (AC) and *B. carinata* (BC), were sequenced. *Bot1* sequences were compared with entries in the GenBank-EMBL database using WU-BLAST2 software (Gish, 2006), and ORFs were identified using 'ORF Finder' from NCBI (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>). Multiple sequence alignment was made using the ClustalW programme (<http://www.ebi.ac.uk/clustalw/index.html>).

Phylogenetic analyses were performed on the set of 16 ORFs of 312 bp isolated in the present study and 13 additional *Bot1* ORF analogues: 6 sequences from *B. rapa* (EMBL accessions nos. AC189341, AC189314, AC189655, AC189480 AC189446 and AC189496), 1 sequence from *B. napus* (AJ245479) and 6 sequences from *B. oleracea* (AC183492, AC183493 and AC183495) including 3 BAC sequences from the 'BoB' library (EU642504-EU642506). A nexus file was created from the multiple sequence alignment with Mesquite v. 1.06 (Maddison and Maddison, 2005). We used MODELTEST v. 3.06 (Posada and Crandall, 1998) to select the most likely model of evolution according to the Akaike information criterion (AIC); the selected model of DNA substitution was the GTR model (Rodríguez *et al.*, 1990), with no invariable site and $R_{mat} = \{0.2737 \ 2.1001 \ 0.8179 \ 1.0216 \ 3.4767 \ 1.0000\}$. Using PAUP v. 4.0b10 (Swofford, 2002), three different phylogenetic methods were applied and compared within each other to evaluate the molecular phylogeny of the 29 *Bot1* ORF-like sequences: the maximum parsimony (MP) method, the neighbour-joining (NJ) method and the maximum likelihood (ML) method, taking into account the optimized parameters obtained from MODELTEST when appropriate. For NJ and ML analyses, bootstrapping of 1000 cycles was conducted. All the phylogenetic trees were obtained

with the 50% majority-rule consensus tree method and were drawn using TreeView v. 1.6.6. All trees were treated as unrooted.

BLASTN was used to align Bo6L1-15 coding sequence with *B. oleracea* WGS and *B. rapa* sequences (whole BAC and BES). Only alignments spanning the whole segment from base 80 to base 250 were included, duplicates were removed. The resulting aligned segments (266 from *B. oleracea* and 14 from *B. rapa*) were aligned to each other using MUSCLE 3.6 (Edgar, 2004). From this alignment a 'median network' phylogeny was calculated (Figure 7) using the 'Network' program (Bandelt *et al.*, 1999) and an artificially rooted phylogenetic tree was built using PhyML to compute maximum likelihood from a neighbour-joined seed tree (Figure S3).

***In silico* analyses**

We identified and analysed the genomic structure of full-length *BotI* elements on the three *B. oleracea* BAC sequences BoB028L16, BoB048N13 and BoB029L16 (accession nos. EU642504-EU642506; 'BoB' library from Warwick-HRI – Ryder, Barker, King, unpublished). We searched these BAC sequences for typical CACTA motifs consisting of (i) terminal inverted repeats (TIRs) that terminate in the conserved CACTA motif and (ii) subterminal repeats (TRs) of 10 to 20bp repeated in direct and inverted orientation. We thus subjected each BAC sequence to dot plot analyses using the program Dotter (Sonnhammer and Durbin, 1995) and we looked for exact or degenerated repeats using the REPuter package (Kurtz *et al.*, 2001). We applied the same strategy to *B. rapa* and *B. napus* BAC sequences to identify and characterize full-length *BotI* analogues in these two *Brassica* species. The sequences of the CACTA elements were aligned (Figure 5) using the Mauve progressive alignment algorithm (Darling *et al.*, 2004). The software provides user with a display layout option (termed

colour multiplicity) that highlight homologous regions appearing across all the sequences (denoted as backbone and coloured in mauve), as well as conserved sequences appearing only in a subset of the aligned sequences. Matches are coloured differently based on which sequence they match in.

Estimation of copy numbers

The 'BoB' BAC library developed by Warwick-HRI and exploited in a previous study to characterize the genomic organization of *Brassica* retrotransposons (Alix *et al.*, 2005) was used to evaluate the copy number and genomic distribution of *Bot1* in the *B. oleracea* genome. For BAC library screening, the protocol described by Alix *et al.* (2005) was followed by probing the clone Bo6L1-15 on colony filters representing 75% of the full 'BoB' library (19,584 clones). Film exposure time was 1 day. When scoring the BAC library hits on autoradiographs, all signals stronger than background were scored.

As previously described (Alix *et al.*, 2005), we analysed the full length sequence of *Bot1* and the *SLL3* gene against the *B. oleracea* whole genome shotgun (WGS) sequences (0.5× - 1× genome coverage) available from The Institute for Genomic Research (TIGR - preliminary sequence data obtained from <http://www.tigr.org>), the complete *Arabidopsis* genomic sequence, and against a genomic sample of *B. rapa* genome sequences. *B. rapa* BAC-end sequences (BES), BAC phase II sequences and complete BAC sequences were extracted from Genbank on Nov. 1st 2007. 198,535 BAC-end sequences were extracted, totalling 154.6Mbp, representing approximately 29% coverage of the *B. rapa* genome, and 527 BAC sequences were extracted, totalling 62.2Mbp, representing approximately 12% coverage of the *B. rapa* genome (Johnston *et al.*, 2005). A relatively stringent E value of 1e-10 and a less stringent value of 1e-02

were used in BLASTN searches using WU-BLAST2 software (Gish, 2006). The formula presented in Zhang and Wessler (2004) was used to generate genome copy number estimates from BLASTN results using the number of hits and length of *Bot1* or *SLL3* sequence.

ACKNOWLEDGEMENTS

We particularly thank Catherine Damerval for valuable discussion on the phylogeny of *Bot1*. Part of Dr K. Alix's work was supported by a fellowship under the OECD Co-operative Research Programme: Biological Resource Management for Sustainable Agriculture Systems. Preliminary sequence data from *B. oleracea* were obtained from The Institute for Genomic Research website at <http://www.tigr.org>. Sequencing of *B. oleracea* was funded by the "National Science Foundation".

REFERENCES

- AGI, Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796-815.
- Alix, K., Baurens, F.-C., Paulet, F., Glaszmann, J.-C. and D'Hont, A. (1998) Isolation and characterization of a satellite DNA family in the *Saccharum* complex. *Genome*, **41**, 854-864.
- Alix, K. and Heslop-Harrison, J.S. (2004) The diversity of retroelements in diploid and allotetraploid *Brassica* species. *Plant Mol. Biol.* **54**, 895-909.
- Alix, K., Ryder, C.D., Moore, J., King, G.J. and Heslop-Harrison, J.S. (2005) The genomic organization of retrotransposons in *Brassica oleracea*. *Plant Mol. Biol.* **59**, 839-851.

- Bandelt, H.-J., Forster, P. and Rohl, A. (1999) Median-Joining Networks for Inferring Intraspecific Phylogenies. *Mol. Biol. Evol.* **16**, 37–48.
- Bateman, A.J. (1955) Self-incompatibility systems in angiosperms. III. Cruciferae. *Heredity*, **9**, 52-68.
- Bennetzen, J.L. (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr. Opin. Genet. Dev.* **15**, 621-627.
- Brunner, S., Pea, G. and Rafalski, A. (2005) Origins, genetic organization and transcription of a family of non-autonomous helitron elements in maize. *Plant J.* **43**, 799-810.
- Casacuberta, J.M. and Santiago, N. (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene*, **311**, 1-11.
- Cui, Y., Brugière, N., Jackman, L., Bi, Y.-M. and Rothstein, S.J. (1999) Structural and transcriptional comparative analysis of the *S* locus regions in two self-incompatible *Brassica napus* lines. *Plant Cell*, **11**, 2217-2231.
- Darling, A.C.E., Mau, B., Blattner, F.R. and Perna, N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394-1403.
- DeMarco, R., Venancio, T.M. and Verjovski-Almeida, S. (2006) SmTRC1, a novel *Schistosoma mansoni* DNA transposon, discloses new families of animal and fungi transposons belonging to the CACTA superfamily. *BMC Evol. Biol.* **6**, 89.
- Deragon, J.-M., Landry, B.S., Pélissier, T., Tutois, S., Tourmente, S. and Picard, G. (1994) An analysis of retroposition in plants based on a family of SINEs from *Brassica napus*. *J. Mol. Evol.* **39**, 378-386.

- Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797.
- Eichler, E.E. and Sankoff, D. (2003) Structural dynamics of eukaryotic chromosome evolution. *Science*, **301**, 793-797.
- Fedoroff, N. (2000) Transposons and genome evolution in plants. *Proc. Natl. Acad. Sci. USA*, **97**, 7002-7007.
- Feschotte, C., Jiang, N. and Wessler, S.R. (2002) Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* **3**, 329-341.
- Gaeta, R.T., Pires, J.C., Iniguez-Luy, F., Leon, E. and Osborn, T.C. (2007) Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell*, **19**, 3403 - 3417.
- Gish, W.R. (2006) WU-BLAST archives: <http://blast.wustl.edu/>.
- Grandbastien, M.-A. (1998) Activation of plant transposons under stress conditions. *Trends Plant Sci.* **3**, 181-187.
- Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A. and Wendel, J.F. (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res.* **16**, 1252-1261.
- He, Z.H., Dong, H.T., Dong, J.X., Li, D.B. and Ronald, P.C. (2000) The rice *Rim2* transcript accumulates in response to *Magnaporthe grisea* and its predicted protein product shares similarity with TNP2-like proteins encoded by CACTA transposons. *Mol. Gen. Genet.* **264**, 2-10.
- Heslop-Harrison, J.S. (2000) Comparative genome organization in plants: from sequence and markers to chromatin and chromosomes. *Plant Cell*, **12**, 617-635.
- IHGSC (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-

921.

- Inaba, R. and Nishio, T. (2002) Phylogenetic analysis of Brassicaceae based on the nucleotide sequences of the *S*-locus related gene, *SLR1*. *Theor. Appl. Genet.* **105**, 1159-1165.
- Inagaki, Y., Hisatomi, Y., Suzuki, T., Kasahara, K. and Iida, S. (1994) Isolation of a *Suppressor-mutator/Enhancer*-like transposable element, *Tpn1*, from Japanese morning glories. *Genes Genet. Syst.* **74**, 141-147.
- Itoh, Y., Hasebe, M., Davies, E., Tadeka, J. and Ozeki, Y. (2003) Survival of *Tdc* transposable elements of the *En/Spm* superfamily in the carrot genome. *Mol. Genet. Genomics*, **269**, 49-59.
- Jacobs, G., Dechyeva, D., Menzel, G., Dombrowski, C. and Schmidt, T. (2006) Molecular characterization of *Vulmar1*, a complete *mariner* transposon of sugar beet and diversity of *mariner*- and *En/Spm*-like sequences in the genus *Beta*. *Genome*, **47**, 1192–1201.
- Jiang, N., Bao, Z., Zhang, X., Eddy, S.R. and Wessler, S.R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**, 569-573.
- Johnston, J.S., Pepper, A.E., Hall, A.E., Chen, Z.J., Hodnett, G., Drabek, J., Lopez, R. and Price, H.J. (2005) Evolution of genome size in Brassicaceae. *Ann. Bot.* **95**, 229–235.
- Kapitonov, V.V. and Jurka, J. (2001) Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. USA*, **98**, 8714-8719.
- Kapitonov, V.V. and Jurka, J. (2007) *Helitrons* on a roll: eukaryotic rolling-circle transposons. *Trends Genet.* **23**, 521-529.
- Kawasaki, S. and Nitasaka, E. (2004) Characterization of *Tpn1* family in the Japanese

- morning glory: *En/Spm*-related transposable elements capturing host genes. *Plant Cell Physiol.* **45**, 933-944.
- Kazazian, H.H. (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626-1632.
- Kubis, S.E., Castilho, A.M.M.F., Vershinin, A.V. and Heslop-Harrison, J.S. (2003) Retroelements, transposons and methylation status in the genome of oil palm (*Elaeis guineensis*) and the relationship to somaclonal variation. *Plant Mol. Biol.* **52**, 69-79.
- Kuhn, G.C.S., Sene, F.M., Moreira-Filho, O., Schwarzacher, T. and Heslop-Harrison, J.S. (2008) Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. *Chromosome Res.* **16**, 307-324.
- Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J. and Giegerich, R. (2001) REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **29**, 4633-4642.
- Le, Q.H., Wright, S., Yu, Z. and Bureau, T. (2000) Transposon diversity in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA*, **97**, 7276-7381.
- Leflon, M., Eber, F., Letanneur, J.-C., Chelysheva, L., Coriton, O., Huteau, V. Ryder, C.D., Barker, G., Jenczewski, E. and Chèvre, A.-M. (2006) Pairing and recombination at meiosis of *Brassica rapa* (AA) × *Brassica napus* (AACC) hybrids. *Theor. Appl. Genet.* **113**, 1467-1480.

- Lenoir, A., Cournoyer, B., Warwick, S., Picard, G. and Deragon, J.-M. (1997) Evolution of SINE S1 retroposons in Cruciferae plant species. *Mol. Biol. Evol.* **14**, 934-941.
- Lysak, M.A., Koch, M.A., Pecinka, A. and Schubert, I. (2005) Chromosome triplication found across the tribe *Brassiceae*. *Genome Res.* **15**, 516-525.
- Maddison, W.P. and Maddison, D.R. (2005) Mesquite: A modular system for evolutionary analysis. Version 1.06. <http://mesquiteproject.org>.
- Meyers, B.C., Tingey, S.V. and Morgante, M. (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **11**, 1660-1676.
- Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H. and Kakutani, T. (2001) Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature*, **411**, 212-214.
- Morgante, M. (2006) Plant genome organisation and diversity: the year of the junk! *Curr. Opin. Biotechnol.* **17**, 168-173.
- Morgante, M., Brunner, S., Pea, G., gler, K., zuccolo, A. and Rafalski, A. (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* **37**, 997-1002.
- Nacken, W.K., Piotrowiak, R., Saedler, H. and Sommer, H. (1991) The transposable element *Tam1* from *Antirrhinum majus* shows structural homology to the maize transposon *En/Spm* and has no sequence specificity of insertion. *Mol. Gen. Genet.* **228**, 201-208.
- Nicolas, S., Le Mignon, G., Eber, F., Coriton, O., Monod, H., Clouet, V., Huteau, V. Lostanlen, A., Delourme, R., Chalhoub, B., Ryder, C.D., Chèvre, A.-M. and

- Jenczewski, E. (2007) Homeologous recombination plays a major role in chromosome rearrangements that occur during meiosis of *Brassica napus* haploids. *Genetics*, **175**, 487-503.
- O'Neill, C.M. and Bancroft, I. (2000) Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homoeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. *Plant J.* **23**, 233-243.
- Ozeki, Y., Davies, E. and Takeda, J. (1997) Somatic variation during long-term subculturing of plant cells caused by insertion of a transposable element in a *phenylalanine ammonia-lyase (PAL)* gene. *Mol. Gen. Genet.* **254**, 407-416.
- Peterson, P.A. (1953) A mutable pale green locus in maize. *Genetics*, **45**, 115-133.
- Pereira, A., Cuypers, H., Gierl, A., Sommers, Z.S. and Saedler, H. (1986) Molecular analysis of the *En/Spm* transposable element system of *Zea mays*. *EMBO J.* **5**, 835-841.
- Posada, D. and Crandall, K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817-818.
- Rocarro, M., Li, Y., Masiero, S., Saedler, H. and Sommer, H. (2005) ROSINA (RSI), a novel protein with DNA-binding capacity, acts during floral organ development in *Antirrhinum majus*. *Plant J.* **43**, 238-250.
- Rodríguez, F.J., Oliver, J.L., Marín, A. and Medina, J.R. (1990) The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**, 485-501.
- Schwarzacher, T. and Heslop-Harrison, J.S.(P.). (2000) Practical *in situ* hybridization. BIOS Scientific Publishers Ltd, Oxford, UK.
- Shiba, H., Takayama, S., Iwano, M., Shimosato, H., Funato, M., Nakagawa, T., Che, F.-S., Suzuki, G., Watanabe, M., Hinata, K. and Isogai, A. (2001) A pollen coat

- protein, SP11/SCR, determines the pollen S-specificity in the self-incompatibility of *Brassica* species. *Plant Physiol.* **125**, 2095-2103.
- Snowden, K.C. and Napoli, C.A. (1998) *Ps1*: a novel *Spm*-like transposable element from *Petunia hybrida*. *Plant J.* **14**, 43-54.
- Snowdon, R.J., Köhler, W., Friedt, W. and Köhler, A. (1997) Genomic in situ hybridization in *Brassica* amphiploids and interspecific hybrids. *Theor. Appl. Genet.* **95**, 1320-1324.
- Sonnhammer, E.L.L. and Durbin, R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1-GC10.
- Strachan, T., Webb, D. and Dover, G. (1985) Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*. *EMBO J.* **4**, 1701-1708.
- Suzuki, G., Kai, N., Hirose, T., Fukui, K., Nishio, T., Takayama, S., Isogai, A., Watanabe, M. and Hinata, K. (1999) Genomic organization of the *S* locus: identification and characterization of genes in *SLG/SRK* region of *S⁹* haplotype of *Brassica campestris* (syn. *rapa*). *Genetics*, **153**, 391-400.
- Swofford, D.L. (2002) *PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods)*; Version 4. Sinauer associates, Sunderland, MA.
- Talbert, L.E. and Chandler, V.L. (1988) Characterization of a highly conserved sequence related to mutator transposable elements in maize. *Mol. Biol. Evol.* **5**, 519-529.
- Town, C.D., Cheung, F., Maiti, R., Crabtree, J., Haas, B.J., Wortman, J.R., Hine, E.E., Althoff, R., Arbogast, T.S., Tallon, L.J., Vigouroux, M., Trick, M. and Bancroft, I. (2006) Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana*

- reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell*, **18**, 1348-1359.
- Turcotte, K., Srinivasan, S. and Bureau, T.E. (2001) Survey of transposable elements from rice genomic sequences. *Plant J.* **25**, 169-179.
- U, N. (1935) Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Jpn. J. Bot.* **7**, 389-452.
- Vicient, C.M., Jääskeläinen, M.J., Kalendar, R. and Schulman, A.H. (2001) Active retrotransposons are a common feature of grass genomes. *Plant Physiol.* **125**, 1283-1292.
- Walbot, V. and Petrov, D.A. (2001) Gene galaxies in the maize genome. *Proc. Natl. Acad. Sci. USA*, **98**, 8163-8164.
- Wang, G.-D., Tian, P.-F., Cheng, Z.-K., Wu, G., Jiang, J.-M., Li, D.-B., Li, Q. and He, Z.-H. (2003) Genomic characterization of *Rim2* / *Hipa* elements reveals a CACTA-like transposon superfamily with unique features in the rice genome. *Mol. Gen. Genomics*, **270**, 234-242.
- Wendel, J.F. (2000) Genome evolution in polyploids. *Plant Mol. Biol.* **42**, 225-249.
- Wicker, T., Guyot, R., Yahiaoui, N. and Keller, B. (2003) CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. *Plant Physiol.* **132**, 52-63.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P. and Schulman, A.H. (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973-982.

- Yang, T.-J., Kwon, S.-J., Choi, B.-S., Kim, J.S., Jin, M., Lim, K.-B., Park, J.Y., Kim, J.-A., Lim, M.-H., Kim, H.-I., Lee, H.-J., Lim, Y.P., Paterson, A.H. and Park, B.-S. (2007) Characterization of terminal-repeat retrotransposon in miniature (TRIM) in *Brassica* relatives. *Theor. Appl. Genet.* **114**, 627-636.
- Yang, Y.W., Lai, K.N., Tai, P.Y. and Li, W.H. (1999) Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J. Mol. Evol.* **48**, 597-604.
- Yu, Z., Wright, S.I. and Bureau, T.E. (2000) *Mutator*-like elements in *Arabidopsis thaliana*: Structure, diversity and evolution. *Genetics*, **156**, 2019-2031.
- Zabala, G. and Vodkin, L.O. (2005) The *wp* mutation of *Glycine max* carries a gene-fragment-rich transposon of the CACTA superfamily. *Plant Cell*, **17**, 2619-2632.
- Zabala, G. and Vodkin, L. (2007) Novel exon combinations generated by alternative splicing of gene fragments mobilized by a CACTA transposon in *Glycine max*. *BMC Plant Biol.* **7**, 38.
- Zhang, X. and Wessler, S.R. (2004) Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. *Proc. Natl. Acad. Sci. USA*, **101**, 5589-5594.
- Zhang, X. and Wessler, S.R. (2005) *BoS*: a large and diverse family of short interspersed elements (SINEs) in *Brassica oleracea*. *J. Mol. Evol.* **60**, 677-687.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article on line:

Figure S1. ClustalW multiple sequence alignment of the 16 *BotI* sequences isolated in the present study (EMBL accessions: AM888354-AM888369). The *BotI* ORF is

highlighted in green; the deletions of respectively 29 and 69bp mentioned in the text are indicated in yellow. The degenerate PCR primers used for amplification are shown in grey.

Figure S2. *BotI-2* used as a query sequence aligned by BLASTN with E-value threshold 0.01 to *B. oleracea* WGS and *B. rapa* BES. The frequency of subject sequence segment alignments overlapping each 100bp segment of the query sequence is represented.

Figure S3. A phylogeny of A and C genome copies of the *BotI* (or *BotI*-like) transposase constructed using PhyML. *B. oleracea* copies are in blue and *B. rapa* in red. The *B. rapa* copies are almost monophyletic.

Table 1. Accessions of the different diploid and allotetraploid *Brassica* species surveyed in the present study

<i>accession code</i>	<i>species (genomes)</i>	<i>2n</i>	<i>subspecies</i>	<i>variety</i>	<i>common name</i>	<i>cultivar/ accession</i>
1*	<i>B. oleracea</i> (CC)	18	<i>oleracea</i>	<i>ramosa</i>	thousand head kale	FO 44-26 ^a
3*			<i>acephala</i>	<i>acephala</i>	kale	FO 49-10 ^a
4			<i>capitata</i>	<i>capitata</i>	cabbage	PO 71-01
5				<i>sabauda</i>	Savoy cabbage	PO 56-20 ^a
6*			<i>botrytis</i>	<i>botrytis</i>	cauliflower	FLE 62-13 ^a
8*			<i>gemmifera</i>		Brussels sprout	BR 62-06 ^a
9			<i>gongyloides</i>		kohl rabi	CRA 21-0 ^a
26			<i>alboglabra</i>		Chinese kale	A12DHd
11	<i>B. rapa</i> (AA)	20	<i>rapifera</i>		field mustard	Chicon ^b
12			<i>oleifera</i>			R 500 ^b
13	<i>B. nigra</i> (BB)	16			black mustard	Junius ^b
14*	<i>B. napus</i> (AACC)	38	<i>oleifera</i>		oilseed rape - Winter	Darmor ^b
18			<i>oleifera</i>		oilseed rape - Spring	Westar ^b
20			<i>oleifera</i>		oilseed rape - Spring	Drakkar ^b
21*			<i>oleifera</i>		oilseed rape - Spring	Yudal ^b
27			<i>oleifera</i>		oilseed rape - ?	-
23	<i>B. juncea</i> (AABB)	36			brown mustard	Picra ^b
24					brown mustard	Varuna ^b
25*	<i>B. carinata</i> (BBCC)	34			Abyssinian mustard	Awassa 67 ^b

* Accessions from which 1kb-fragments of *Bot1* were isolated.

Seeds were kindly provided by ^a Pr G. Thomas and ^b Dr A.-M. Chèvre, UMR APBV, Rennes, France.

Table 2. Main characteristics of the *Bot1*-like CACTA elements identified in *B. rapa* and *B. napus* BAC clones mined from the database

<i>BAC ID</i>	<i>Position within BAC</i>		<i>size</i>	<i>TIR sequence</i>
	<i>start</i>	<i>stop</i>		
<i>B. rapa</i>				
gi 110796994 gb AC189314.1	21685	39349	7664	CACTACAAGAAAACA
gi 110797021 gb AC189341.1	32642	40439	7797	CACTACAAGAAAACA
gi 110797126 gb AC189446.1	5465	13321	7856	CACTACAAGAAAACA
gi 110797160 gb AC189480.1	33966	43353	9388	CACTACAAGAAAACA
gi 110797176 gb AC189496.1	56852	64624	7773	CACTACAAGAAAACA
gi 110797335 gb AC189655.1	61104	68093	6990	CACTACAAGAAAACA
<i>B. napus</i>				
gi 7657870 emb AJ245479.1	11308	19465	8158	CACTACAAGAAAACAGC

Table 3. Summary of results obtained by BLASTN comparison at two different E values of *Bot1* and the *SLL3* gene to 1) the 0.5× coverage set of *B. oleracea* whole genome shotgun sequences from TIGR and to 2) the set of *B. rapa* genome BAC-end, BAC phase II and complete BAC sequences from Genbank.

For the analysis against *B. rapa* sequences, two estimates are given, in each case the first estimate being derived from alignments to BAC-ends, and the second to whole BAC sequences. The formula for estimating genome copy number is likely to be more reliable for the BAC-end estimates, since the BAC-ends are pseudo-randomly distributed throughout the genome whereas the whole BACs had been deliberately chosen to form a tiling path.

<i>Query sequence</i>	<i>BLASTn hits at 1.e-02</i>	<i>Estimate of genome copy number at 1.e-02*</i>	<i>BLASTn hits at 1.e-10</i>	<i>Estimate of genome copy number at 1.e-10*</i>
1) against <i>B. oleracea</i> WGS				
Bo6L1-15: full 1010bp	1,366	1,605	1,209	1,420
ORF 312bp	757	1,534	719	1,457
Bo1L1-04: full 991bp	1,371	1,629	1,213	1,441
ORF 312bp	758	1,536	725	1,470
Bo8L1-11: full 994bp	1,367	1,621	1,215	1,441
<i>SLL3</i> : genomic 2067bp	5,298	3,802	3,463	2,485
coding 1488bp	4,646	4,237	3,434	3,132

2) against <i>B. rapa</i> BES and BAC sequences				
Bo6L1-15: full 1010bp	30	15 - 167	28	13 - 160
ORF 312bp	28	22 - 161	28	22 - 161
Bo1L1-04: full 991bp	30	15 - 169	28	14 - 160
ORF 312bp	28	22 - 161	28	22 - 161
Bo8L1-11: full 994bp	31	15 - 177	28	14 - 160
<i>SLL3</i> : genomic 2067bp	163	51 - 910	70	21 - 393
coding 1488bp	143	56 - 797	70	27 - 394

* Calculated taking into account length of sequences being compared, probability that 1 hit = 1 element and relative genome coverage (see Zhang and Wessler 2004).

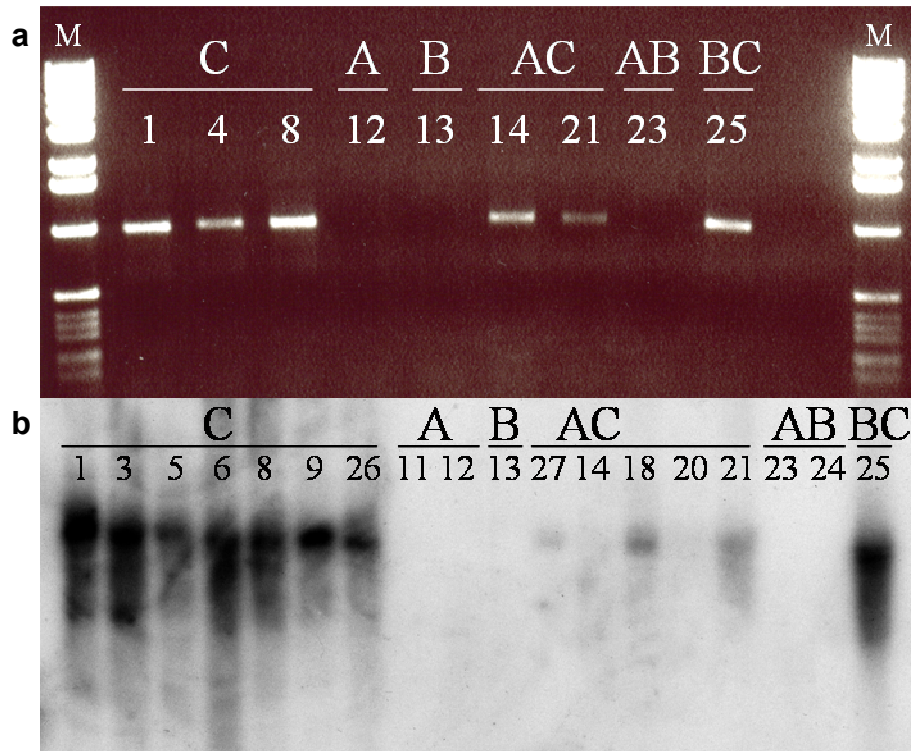


Figure 1. Isolation of a *Brassica* C genome-specific sequence. **(a)** PCR amplification of a 1kb-band from genomic DNA using the single primer BEL1MF shown by gel electrophoresis. The 1kb- band is specifically amplified in *Brassica* species that possess the C genome. **(b)** *Eco*RI digests of genomic DNAs of the three diploid and three allotetraploid *Brassica* species of U's triangle after Southern hybridization with the *Bot1* clone Bo6L1-15. Only the C genome species show hybridization signals demonstrating the C genome-specificity of the probe. Variations of hybridization signals among the different *B. napus* accessions are mainly due to variations of DNA loading (seen after gel staining with ethidium bromide; data not shown).

Genomes are designated; see Table 1 for accession identifications. Molecular weight marker (M): 1kb ladder.

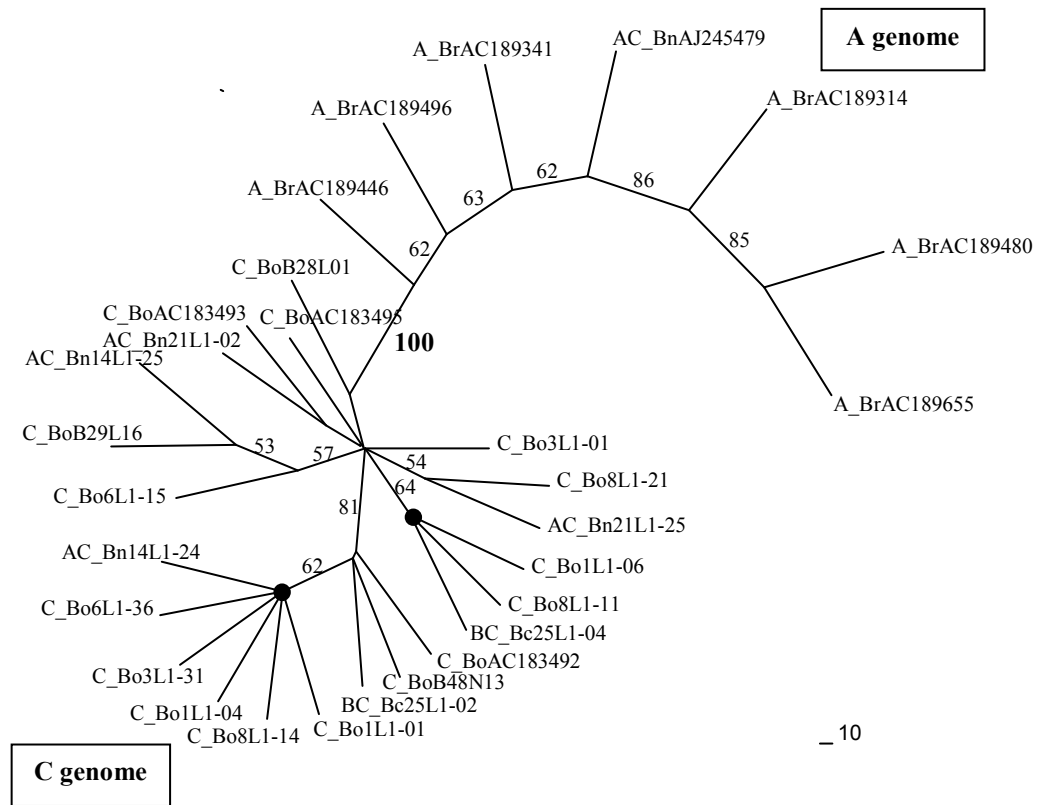


Figure 2. An unrooted maximum likelihood consensus tree of *BotI* based on the 312bp-long ORFs of the 16 sequences of 1kb isolated in the present study and of 13 additional analogous sequences obtained from BAC sequencing or mined from the database. Numbers indicate the percentage values (≥ 50) from 1000 bootstrap replicates supporting a particular clade. Polytomies are shown by (●). Genome composition for each accession is indicated: A for *B. rapa* (Br), C for *B. oleracea* (Bo), AC for *B. napus* (Bn) and BC for *B. carinata* (Bc).

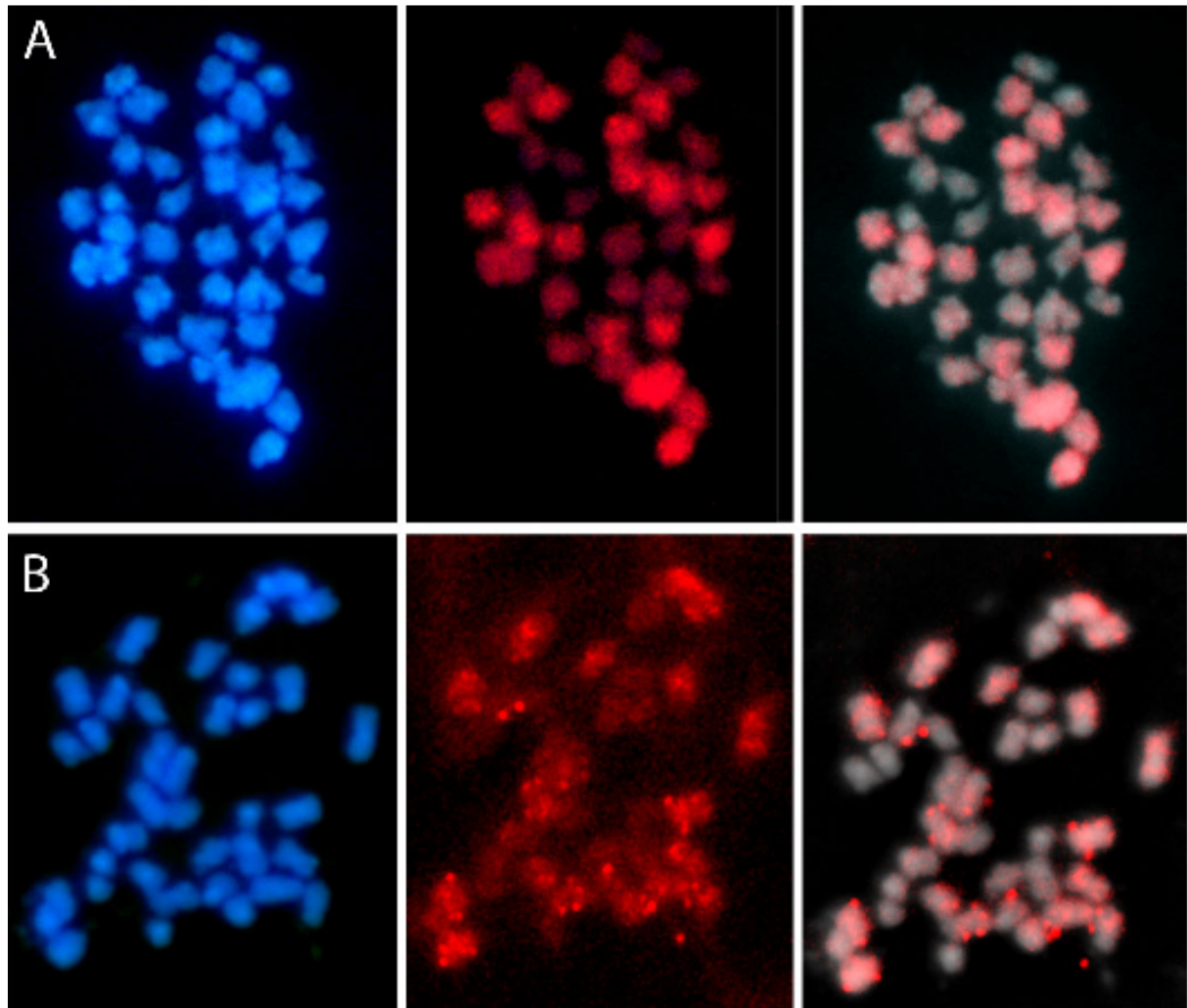


Figure 3. Fluorescent *in situ* hybridization of *BotI* (clone Bo6L1-15) to metaphase chromosomes of *B. napus* cv. 'Yudal' (A) and 'Drakkar' (B). Chromosomes are counterstained blue with DAPI (left), and probe hybridization sites are shown in red (centre), and in overlay (right). *BotI* labels the eighteen C-genome chromosomes in each metaphase originating from *B. oleracea* strongly and relatively uniformly along their lengths, with some stronger and weaker areas around centromeres. Only weak hybridization is seen to the 20 A genome chromosomes from *B. rapa* (probe in Yudal labelled with digoxigenin detected with antidigoxigenin-FITC and colour converted to red; probe in Drakkar labelled with biotin detected with streptavidin-Alexa 594 where large red dots are not considered signal). Scale bar = 5 μ m.

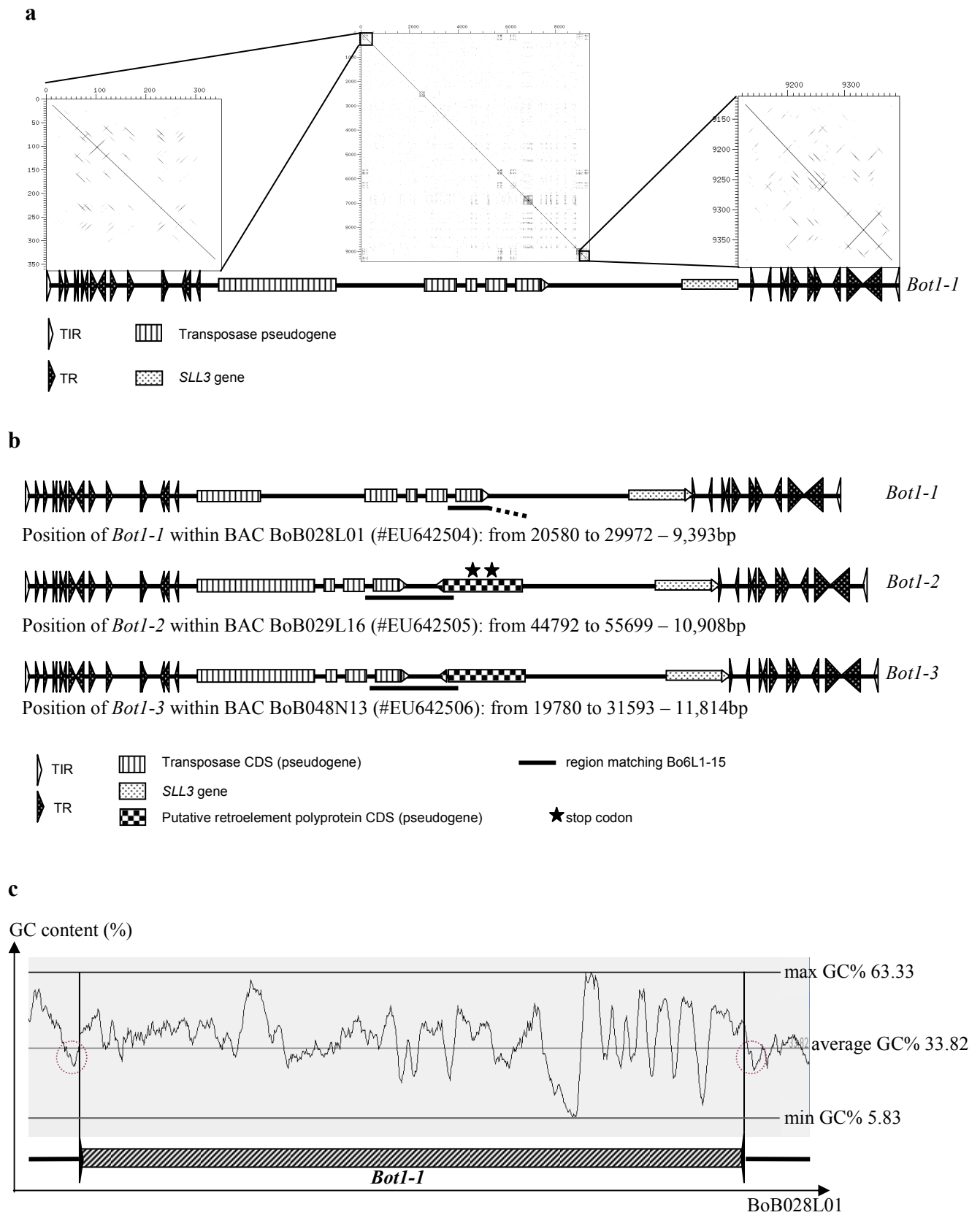


Figure 4. Schematic representation of the overall genomic structure of the different *B.*

oleracea CACTA transposons studied. **a.** Dot-plot pattern and corresponding overall structure of *Bot1-1* identified in the *B. oleracea* BAC BoB028L01. **b.** Structural comparisons of the CACTA transposons. **c.** GC content of *Bot1-1* and its flanking regions which correspond to AT-rich regions (encircled on the figure). The median line corresponds to the average GC content of the entire BAC sequence analysed (window size 120).

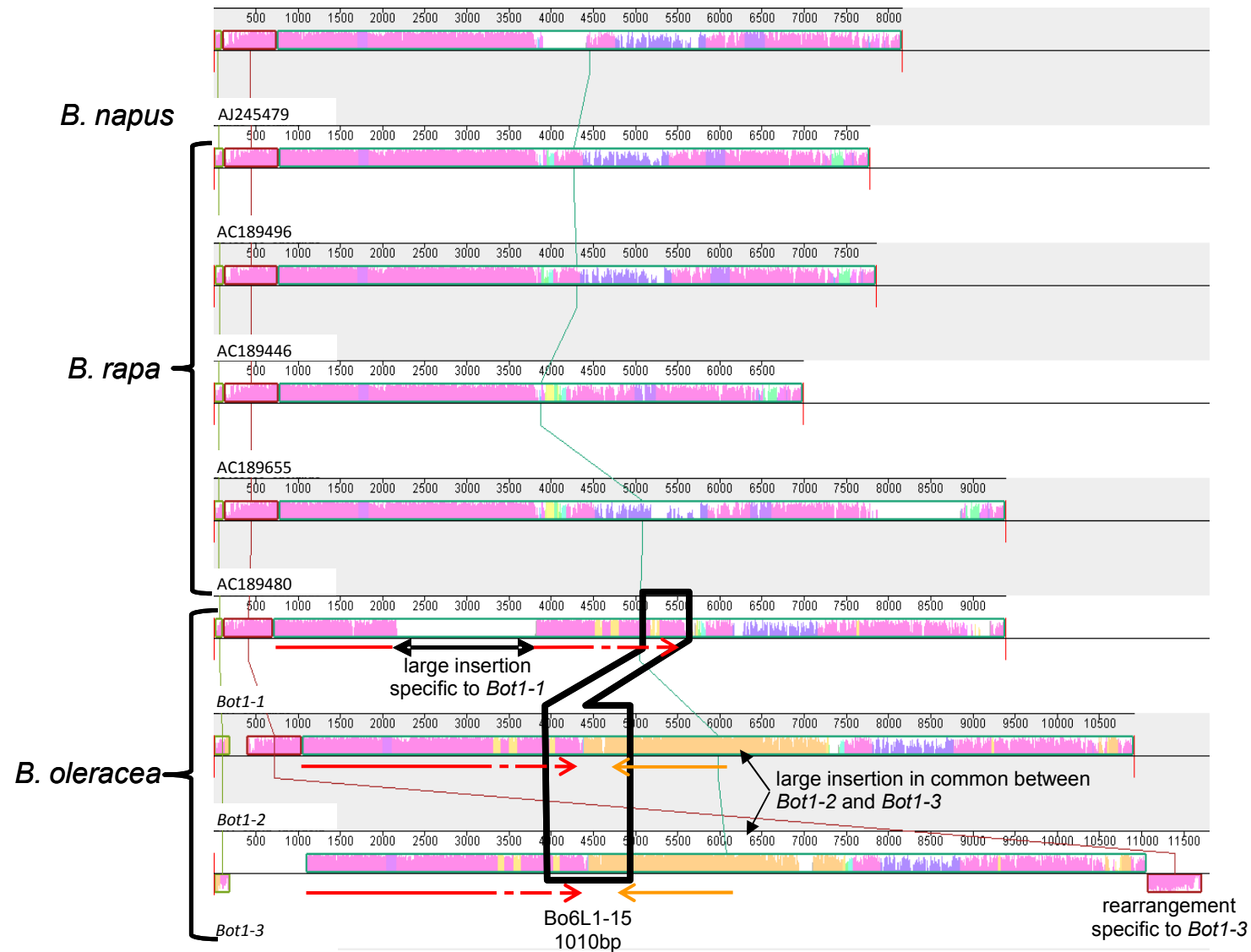


Figure 5. Multiple alignment of *Bot1* elements from *B. napus* (AJ245479) *B. rapa* (AC189496, AC189446, AC189655, AC189480) and *B. oleracea* (*Bot1-1*, *Bot1-2*, *Bot1-3* with BAC accession numbers). A similarity profile using Mauve multiple alignment software (Darling *et al.* 2004) is shown

for each sequence. Regions exhibiting no homology with any other sequence are uncoloured, e.g. the large insertion annotated in *Bot1-1*. Homologous regions appearing across all the sequences (denoted as backbone) are coloured mauve. Conserved regions appearing only in a subset of the aligned sequences are coloured differently based on which sequence they match. Orange regions delineate fragments found only in *B. oleracea* which appear C genome-specific. The region homologous to the B6L1-15 sequence is shown in the black box. (→) transposase pseudogene CDS (→) retroelement polyprotein pseudogene CDS, as described in Figure 4.

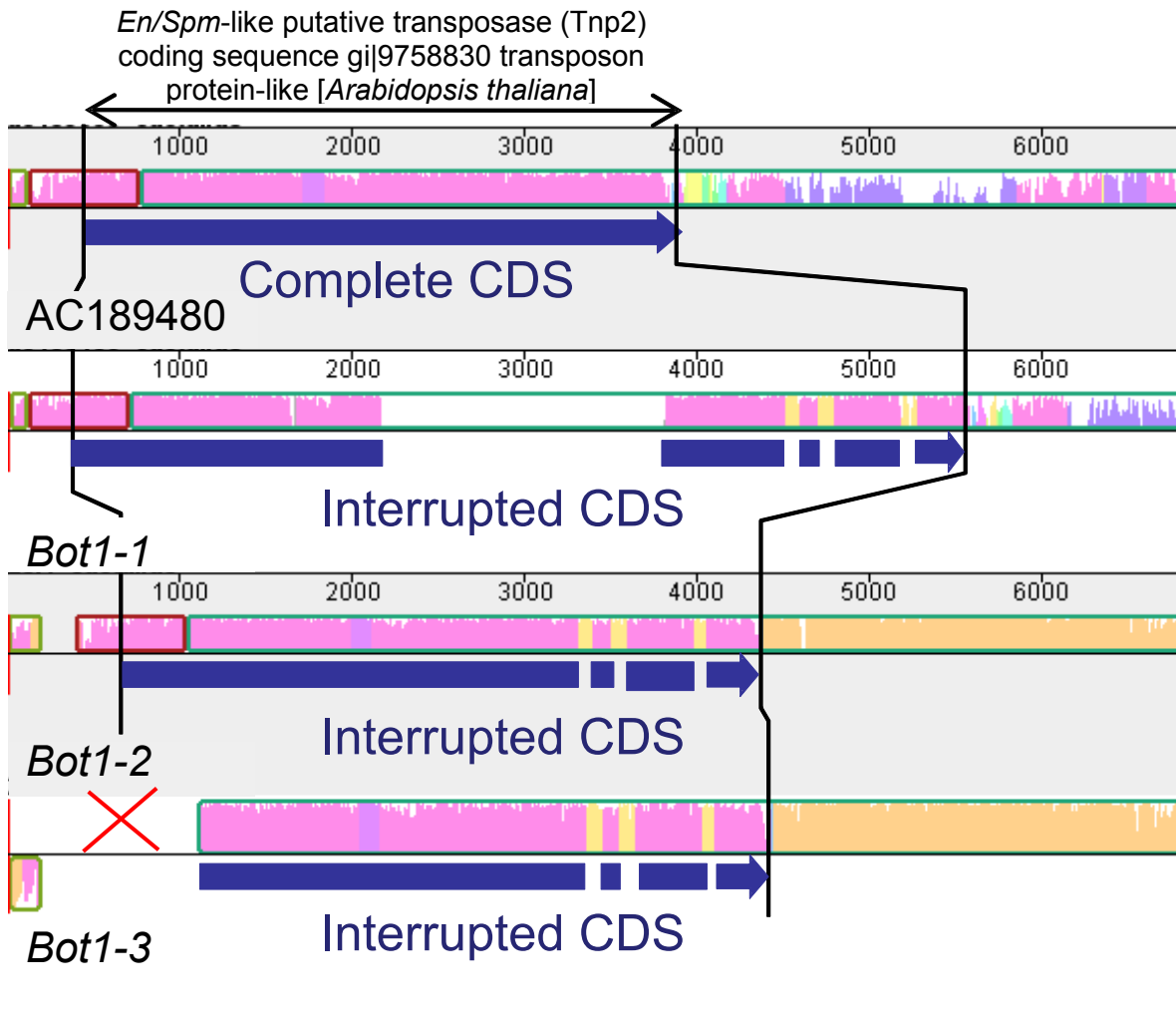


Figure 6. Multiple alignment analysis reveals that *BotI* elements of *B. oleracea* contain an interrupted version of the *En/Spm*-like transposase gene. The blue arrow corresponds to the transposase CDS. A full CDS is present in *B. rapa* *BotI*-like element. The *En/Spm*-like gene is interrupted by fragments of putative C genome specific DNA in *B. oleracea* *BotI* CACTA transposable elements.

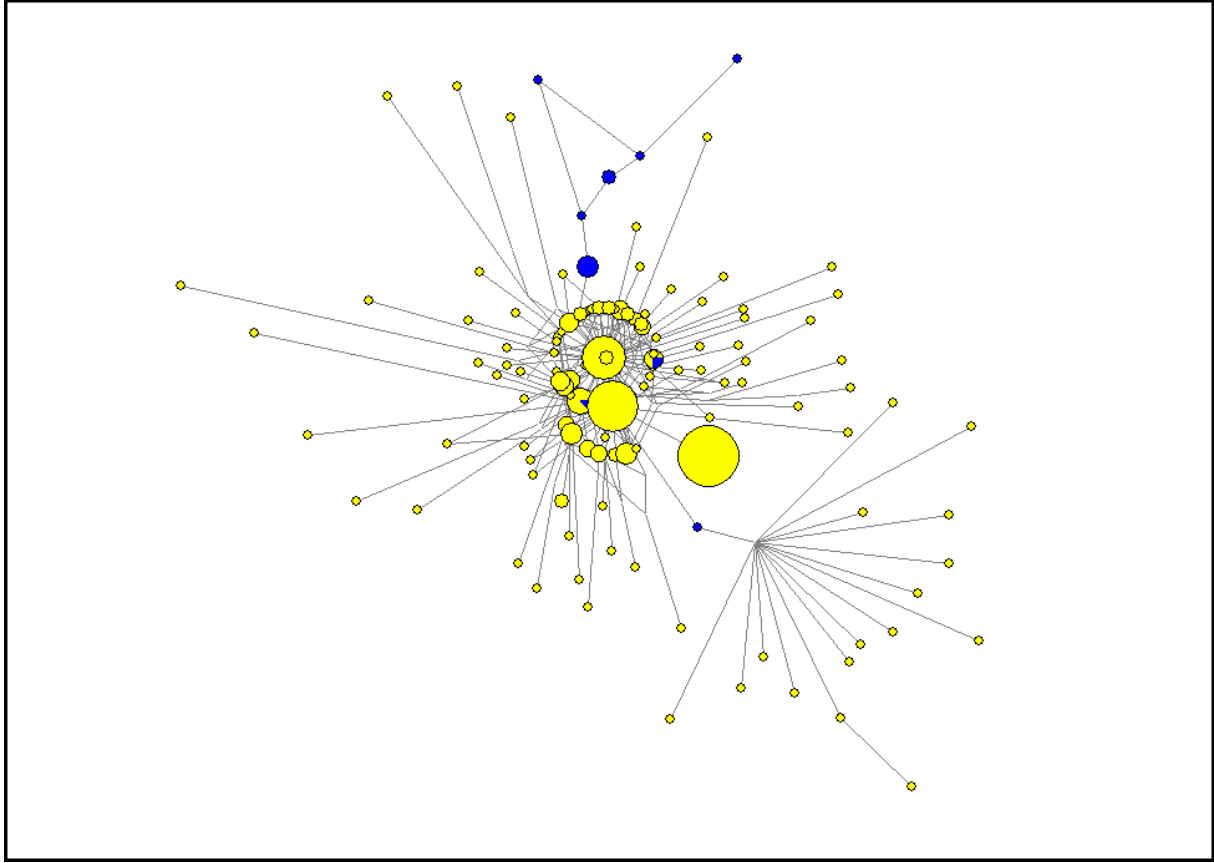


Figure 7. A median network phylogeny of aligned segments of the *B. oleracea* WGS (yellow) and *B. rapa* BAC sequences (blue) to the transposase segment of *Bot1*. The smallest circles are single copies of the transposon. The larger circles represent 'star contractions', that is, they represent multiple copies which appear to have expanded from a single common ancestral sequence, from which they differ by 5 or less mutations.